A proposed discrete distribution for the statistical modeling of Likert data

Kidd, Martin
Centre for Statistical Consultation
University of Stellenbosch, Private Bax X1
Matieland 7602, South Africa
E-mail: mkidd@sun.ac.za

Laubscher, Nico
InduStat Pro
Stellenbosch, 7600, South Africa
E-mail: nfl@industat.co.za

Abstract

When Likert scale data are subjected to statistical analyses, the normal distribution is usually assumed as underlying distribution. Alternatively nonparametric statistical techniques are applied. Other techniques like polychoric correlation assumes that the Likert scale divides the sample space of the normal distribution into intervals. In this paper, an alternative distribution based on the normal distribution is proposed. The sample space is assumed to be discrete and consists only of the values of the Likert scale. This distribution has two parameters (one for location and one for scale) corresponding to those of its normal counterpart. This (what will be called the Likert) distribution differs from the normal distribution in that its shape depends on both parameters.

A numerical procedure for obtaining maximum likelihood estimators for the two parameters is exhibited and some desirable properties of the distribution discussed. There are theoretical aspects of the distribution that remain to be researched and the purpose of this paper is to present the initial concept and to test its acceptability among peers.

Results from a study on real world Likert scale data indicate that in 67% of goodness-of-fit tests, the Likert distribution provided an acceptable fit at a 5% significance level.

A test statistic based on the Likert distribution is proposed for comparing means of two groups, and results from a comprehensive simulation study indicated superior power of this test over the standard t-test for small samples.

1. Introduction

The Likert scale is widely used for measuring latent variables through the use of questionnaires. It takes on discrete specified ordinal values eg 1, 2, 3, 4, 5, and in many cases descriptive words like "*Completely Disagree*" to "*Completely Agree*" accompany such a scale.

Statistical analyses of Likert scale data take on many forms from comparing different groups, doing correlation analyses, to more complex analyses like factor analysis and structural equations modeling. In most of these cases the data are assumed to come from a normal distribution, or where appropriate nonparametric techniques are applied. Other techniques like tetrachoric and polychoric correlation assume that the Likert scale divides the sample space of the normal distribution into intervals, and then the statistical techniques are derived from this assumption.

In this paper a different distribution based on a discrete sample space defined by the Likert scale is introduced. The basic concepts of the distribution are presented in section 2. Sections 3 and 4 deal with the expected value and maximum likelihood estimators for the parameters. In section 5 goodness-of-fit tests done on real world data are reported to give an indication of the appropriateness of this proposed distribution. A test statistic for comparing the means of two groups is proposed in section 6. A summary and outline of future work are presented in section 7.

2. The Likert Distribution

The sample space of the proposed distribution is a discrete ordinal sample space taking on the values of the Likert scale. For example, for a 5-point Likert scale, the sample space typically consists of the integers 1, 2, 3, 4, 5. Thus the sample space is an ordered set of consecutive integers. What will be referred to as the Likert distribution, then assigns probabilities to each of the sample points based on two parameters, μ and σ similar to the parameters of a normal distribution. The proposed probability mass function for the distribution based on a sample space of contiguous integer-valued points $S = \{k_1, k_1+1, \ldots k_2-1, k_2\}$ is defined as:

$$f(x|\mu,\sigma) = \frac{1}{K(\mu,\sigma)} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

where $x \in S$, $\mu \in (-\infty, +\infty)$, $\sigma \in (0, +\infty)$ and

$$K(\mu,\sigma) = \sum_{j=k_1}^{k_2} e^{-\frac{1}{2}\left(\frac{j-\mu}{\sigma}\right)^2} .$$

The expression $K(\mu,\sigma)$ ensures that $f(x|\mu,\sigma)$ is a probability function.

Some noteworthy properties of the distribution are the following:

1. The larger the difference between x and μ, the smaller the point probability $f(x|\mu,\sigma)$.
2. As $k_1 \to -\infty$ and $k_2 \to +\infty$ then

   $\sum_{j=k_1}^{k_2} e^{-\frac{1}{2}\left(\frac{j-\mu}{\sigma}\right)^2} \to \sqrt{2\pi\sigma^2}$ and thus the distribution tends to the normal distribution. This property was numerically verified, but still requires theoretical proof.

3. As $\sigma \to +\infty$, then $\sum_{j=k_1}^{k_2} e^{-\frac{1}{2}\left(\frac{j-\mu}{\sigma}\right)^2} \to k_2 - k_1 + 1$ and $f\left(x\right) = \frac{1}{k_2 - k_1 + 1}$, the uniform distribution.
4. As $\mu \to +\infty$, then $f(k_2) \to 1$ and as $\mu \to -\infty$, then $f(k_1) \to 1$
5. The shape of the distribution depends on both μ and σ. When μ=middle value of the Likert scale, then the distribution is symmetric. As $\mu \to +\infty$, the distribution becomes left skewed and as $\mu \to -\infty$, it becomes right skewed. Increasing σ flattens out the distribution until it eventually becomes a uniform distribution (see point 3).

3. Expected value of the distribution.

The expected value of the distribution is given by:

$$E\left[x|\mu,\sigma\right] = \sum_{j=k_1}^{k_2} j \cdot f(j|\mu,\sigma)$$

$$= \frac{1}{K\left(\mu,\sigma\right)} \cdot \sum_{j=k_1}^{k_2} j \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

It is important to note here is that μ is not the expected value of the distribution. The expected value lies between k_1 and k_2, whereas μ can range between $-\infty$ and $+\infty$. As $\mu \to +\infty$, $E\left[x\right] \to k_2$ and as $\mu \to -\infty$, $E\left[x\right] \to k_1$ (see point 4 in section 2).

4. Maximum Likelihood Estimation

For a set of realisations of *x* under the Likert distribution, say, $x_1, \ldots, x_n$, let:

$$K = K\left(\mu,\sigma\right)$$

$$u_i = \frac{x_i - \mu}{\sigma}$$

$$v_j = \frac{j-\mu}{\sigma}$$

$$w_j = e^{-\frac{1}{2}v_j^2}.$$

Then the likelihood function is:

$$LF = \prod_{i=1}^{n} K\left(\mu,\sigma\right)^{-1} \cdot e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2}$$

$$= K^{-n} \cdot \prod_{i=1}^{n} e^{-\frac{1}{2}u_i^2}$$

From this the log likelihood can be written as:

$$\ln LF = -n \ln K - \frac{1}{2}\sum_{i=1}^{n} u_i^2 .$$

To estimate μ and σ, the above expression is maximised wrt μ and σ.

The derivatives with respect to μ and σ can be written as:

$$\frac{\partial}{\partial \mu}\ln LF = \frac{1}{\sigma}\left[\sum_{i=1}^{n} u_i - \frac{n}{K}\sum_{j=k_1}^{k_2} v_j w_j\right] \text{ and}$$

$$\frac{\partial}{\partial \sigma}\ln LF = \frac{1}{\sigma}\left[\sum_{i=1}^{n} u_i^2 - \frac{n}{K}\sum_{j=k_1}^{k_2} v_j^2 w_j\right].$$

Numerical algorithms can be used to solve for μ and σ from the above ML equations. The solution will be denoted by $\hat{\mu}$ and $\hat{\sigma}$ respectively.

Of course, if $\hat{\mu}$ and $\hat{\sigma}$ are the MLE's of μ and σ, then $\hat{E}_L \equiv E\left(x \middle| \hat{\mu}, \hat{\sigma}\right)$ will be the MLE of the expected value. A property empirically observed was that $\hat{E}_L = \frac{1}{n}\sum_{i=1}^{n} x_i$. This means that the sample arithmetic mean equals the MLE of the expected value of the Likert distribution.

5. Goodness-of-fit on actual data

To get an idea of how well the Likert distribution fits actual data, 697 data sets were used, and tests done to check whether the distribution fits the data. No claim is made that this collection of data sets is a representative sample from the population of all real world data sets, but it does give an indication of the validity of the distribution. The following results emerged:

- On a 5% significance level, 33% of the data sets did not support the Likert hypothesis (the null-hypothesis was rejected by the goodness-of-fit test). This means that 67% of the data sets did not contradict the Likert distribution hypothesis.
- For “smaller” sample sizes (n<200) the % rejected dropped to 24%.
- There was a trend that the goodness-of-fit increased for Likert scales with a smaller number of outcomes. For 4-point Likert scale data, only 15% (7% for n < 200) of the tests were rejected. For 7-point scale data, the % rejected increased to 50% (42% for n < 200).

6. Comparing two Likert distribution group means

In order to test for equality of the means of two groups using the Likert as underlying distribution, the following test statistic is proposed:

Let $\left(\hat{\mu}_1, \hat{\sigma}_1\right)$ and $\left(\hat{\mu}_2, \hat{\sigma}_2\right)$ be the maximum likelihood estimates of the Likert parameters obtained from the two random samples, and

$$\bar{x}_1 = E\left[x\middle|\hat{\mu}_1,\hat{\sigma}_1\right] = \frac{1}{n_1}\sum_{i=1}^{n_1} x_{1i}\ , \quad \bar{x}_2 = E\left[x\middle|\hat{\mu}_2,\hat{\sigma}_2\right] = \frac{1}{n_2}\sum_{i=1}^{n_2} x_{2i}\ ,$$

be the Likert expected value MLE's for the two samples sets respectively.

The difference of the sample means, $L = \bar{x}_1 - \bar{x}_2$, is proposed as test statistic for the null-hypothesis that the samples come from two Likert populations with equal expected values.

The distribution of L is determined through simulation by drawing $B(=1000)$ pairs of random samples of sizes n_1 and n_2 from the Likert distribution using parameter sets $(\hat{\mu}_1,\hat{\sigma}_1)$ and $(\hat{\mu}_2,\hat{\sigma}_2)$ respectively. The p-value of the test statistic for the data is then determined from the location of 0 in the simulated empirical distribution.

A comprehensive simulation study was conducted to compare this Likert test with the standard t-test (assuming normality of the data). Various parameters like sample sizes, effect sizes etc were randomly varied in this simulation study. Data was simulated from the Likert distribution.

Results from this study showed that in the majority of cases, the Likert test and t-test gave the same outcomes (both either rejecting or accepting the null hypothesis), especially for "larger" sample sizes. The simulation did however show, that for "small" samples $(n < 20)$, the Likert test was more inclined to indicate significant differences than the t-test. Figure 1 shows an extract of the simulation results where the Likert test was compared to the t-test and a bootstrap test for the equality of two means. The figure indicates that with increasing effect size, the Likert test had superior power over the other two tests.

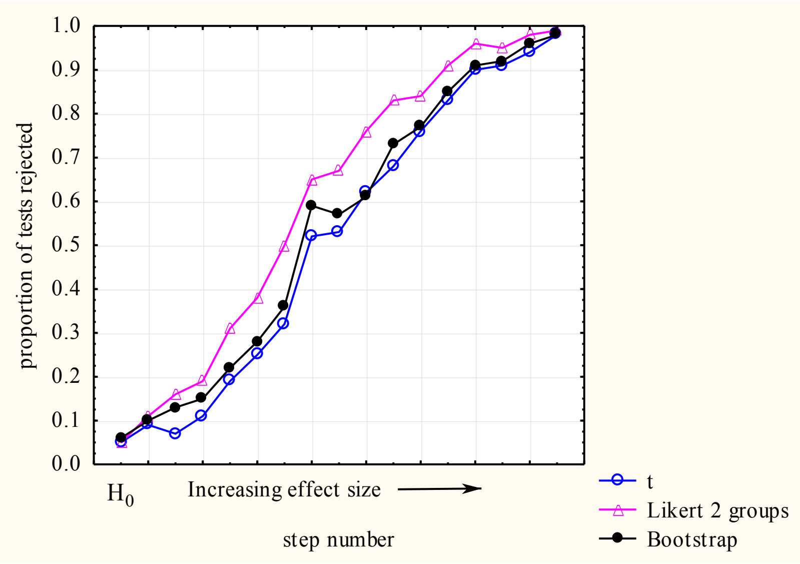

Figure 1 Results from a simulation study indicating superior power of the Likert two groups test over the t-test and bootstrap test for small samples ($n_1 = n_2 = 15$).

7. Summary and further research

This paper proposes a distribution for analysing Likert scale data based on the normal distribution. Desirable properties, linking it to the normal distribution were shown. Some of the properties presented here, have been theoretically derived and others have been numerically verified (still to be proven theoretically).

A test statistic for comparing means of two samples from the Likert distribution was proposed, and simulation studies suggested possible advantages over the standard t-test for small samples.

An important extension of this work will be to extend this distribution to the bivariate case. This should then enable one to calculate correlations based on the Likert distribution. Correlations are important in the analysis of multivariate Likert scale data because factor analysis, structural equations modeling (SEM) etc, are all techniques that are based on covariances and correlations.

REFERENCES

Tamhane, Ajit C, Ankenman, Bruce E, Yang, Ying (2002). *The Beta Distribution as a latent response model for ordinal data (I): Estimation of Location and Dispersion Parameters*. J.Statist. Comput. Simul., 2002, Vol. 72(6), pp. 473-494.

Poon, Wai-Yin (2004). *A latent normal distribution model for analysing ordinal responses with applications in meta-analysis.* Statist. Med. 2004; 23:21552172.

Tang, Man-Lai, Poon, Wai-Yin (2007). *Statistical inference for equivalence trials with ordinal responses: A latent normal distribution approach.* Computational Statistics & Data Analysis 51 (2007) 5918-5926.

Olsson, Ulf (1979). *Maximum likelihood estimation of the polychoric correlation coefficient.* Psychometrika, Vol. 44, No. 4, pp.443-460